Deep Learning for Multi-Instrumental Music Generation

Constance Horng Jimming He Ethan Ko

Department of Computer Science, Stanford University

Predicting

Creative endeavors such as art and music once differentiated man and machine. However, given recent advancements in artificial intelligence and machine learning, this border has become less defined. As musicians ourselves, we wanted to explore the intersection between technology and music. In this project, we trained RNN, 2D CNN, and LSTM models on a cleansed, multi-instrumental subset of the Lakh Pianoroll Dataset (LPD) to generate original track samples of music consisting 5 instruments: piano, drums, guitar, bass, and strings.

Data/Features

We used a cleansed subset of the Lakh Pianoroll Dataset (LPD) comprising of 21,245 MIDI files for 5 different instruments, all standardized to a 4/4 time signature and retaining only the highest confidence score file for each song. We standardized each beat's length using symbolic timing, and we adopted a temporal resolution of 24/beat to capture common temporal patterns. The note pitches encompassed 128 potential notes, spanning from C1 to G9.

RNN Model

- Baseline model
- Chose RNN because it is used for processing sequential data or time series data
- Transformed the multi-instrumental pianoroll data at each time step by flattening it, generating a one-dimensional vector
- Then used the previously learned vectors to predict the vector at the next time-step

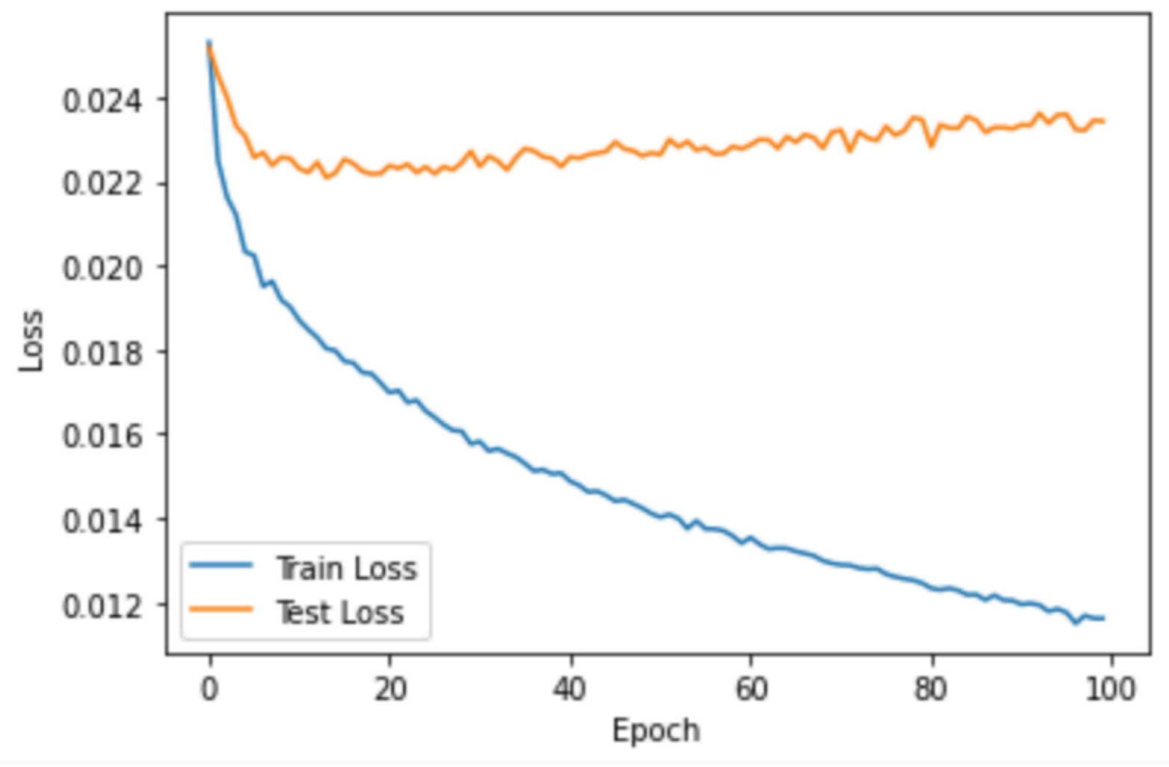

Figure 1. Train and test loss for our baseline RNN model

- Generated music with a considerable amount of dissonance and seemingly random sounds
- Struggled to strike a balance between sampled note quantity per instrument and maintaining the multi-instrumental character of the composition

2D CNN Model

- CNN model with four 2D convolution layers
- Dropout layers of 0.4, batch normalization
- Dependent sampling among the 5 instruments reduced randomness in note pitches and rhythms and created more harmonious music

2D CNN Model

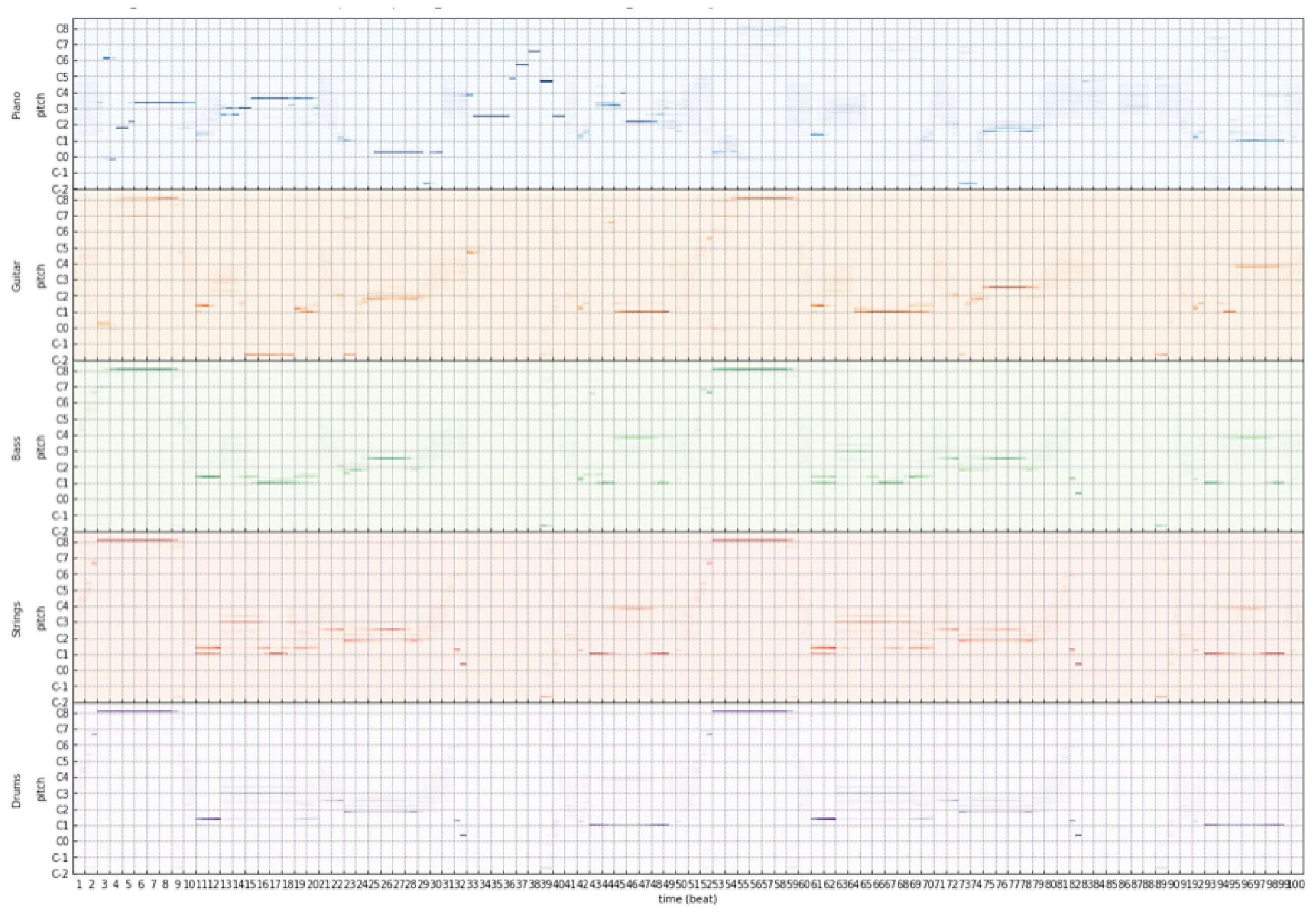

Figure 2. Visualization of the pitch generation at each time step for the 5 instruments

LSTM Network

- LSTMs have built-in mechanisms to mitigate some of the issues seen with RNNs such as vanishing or exploding gradients
- Can also capture long-term dependencies effectively
- 3 LSTM layers, each with 512 units, followed by a fully connected Dense layer with softmax activation for the output
- Dropout layers (0.4 rate) to reduce overfitting
- Fed in the flattened vector representation of the multi-instrumental data at each timestep as input

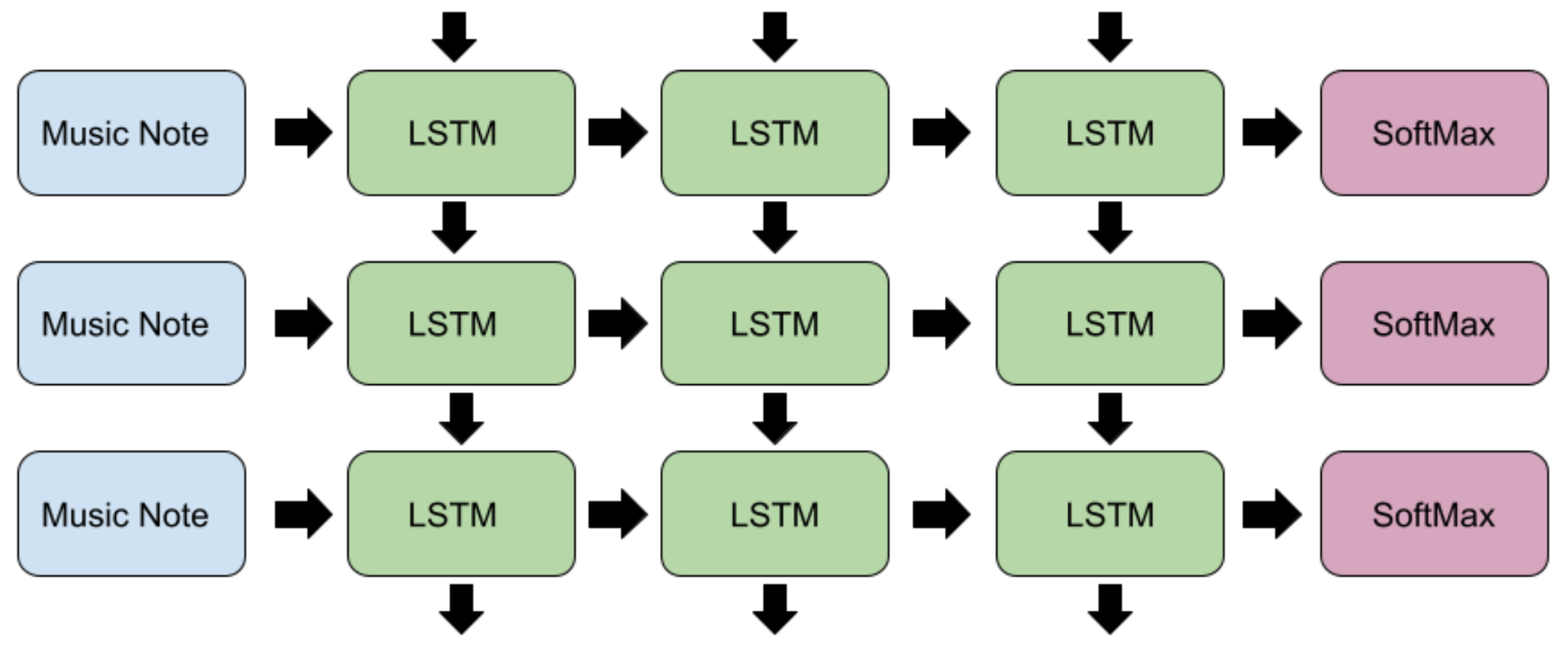

Figure 3. Condensed LSTM Model Architecture

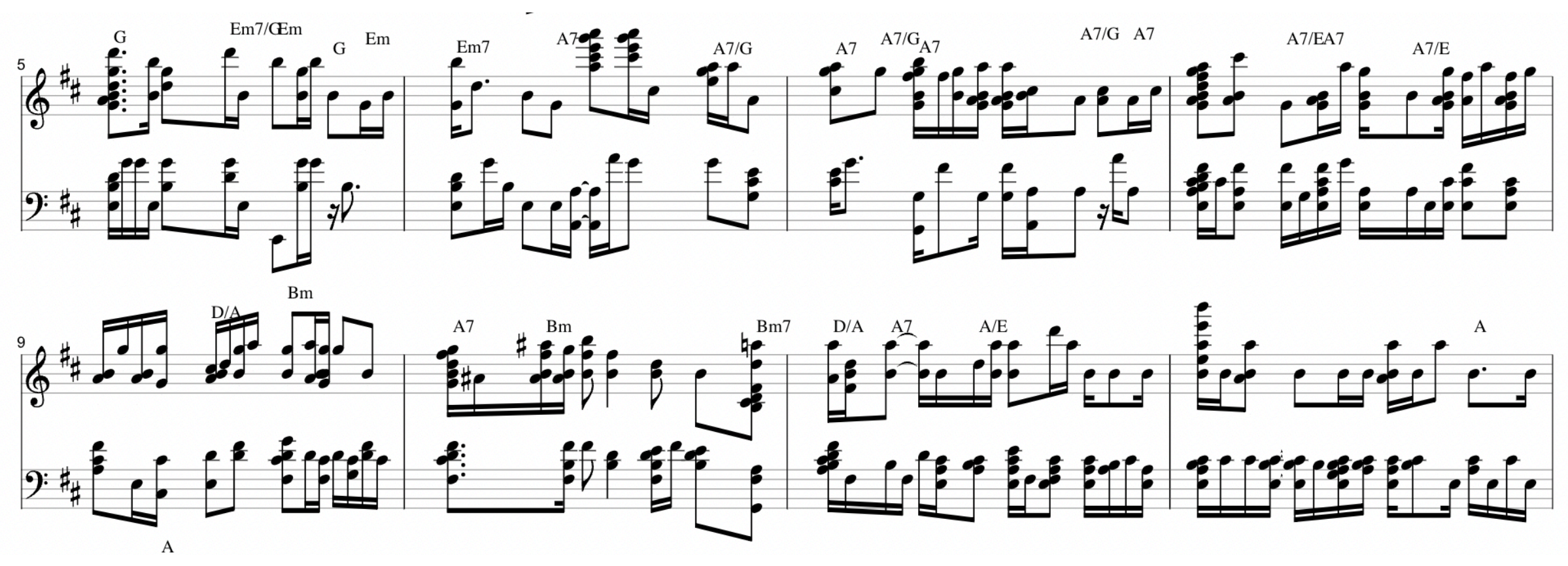

Figure 4. Sheet music for an excerpt from our best generated sample (condensed into 2 staves)

Results

Model	Average Rating
Baseline LPD-5	9.3
Baseline RNN #1	4.9
Baseline RNN #2	4.6
2D CNN #1	5.8
2D CNN #2	6.1
LSTM #1	7.8
LSTM #2	8.3

Table 1. Average Ratings for the Baseline RNN, 2D-CNN, and LSTM models.

Discussion

1. Baseline RNN: failed to maintain a balance between the number of sampled notes per instrument and the multi-instrumental character of the composition, causing dissonance and random-sounding parts
2. 2D CNN: performed better than the baseline RNN model due to its reliance on dependent instrument sampling, leading to more cohesive connections among the different instruments and an overall higher aesthetic quality.
3. LSTM: highest average ratings overall, leveraged the ability of LSTM layers to remember long-term dependencies, alleviating the issue of dissonance and generating sequences that are more harmonious and melodically pleasing, memory retention allowd for emulation of complex chord progressions and rhythms

Our evaluation shows that it can be the LSTM multi-track music generation model can be a good alternative to RNNs and CNNs as the generated music samples produced significantly more aesthetically pleasing music, receiving higher ratings across the board.

Future Work

1. Transformers: attention mechanisms and sequence generation abilities can potentially improve the learning of melodic dependencies to solve our dissonance issues.
2. GANs: can potentially allow for easier extraction of features such as pitch and timbre and generate new music that is similar in quality to the pieces they were trained on.

References

[1] Michael Conner, Lucas Gral, Kevin Adams, David Hunger, Reagan Strelow, and Alexander Neuwirth. Music generation using an lstm, 2022.

[2] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Lakh pianoroll dataset.

[3] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential gans for symbolic music generation and accompaniment, 2017.

[4] Vaishali Ingale, Anush Mohan, Divit Adlakha, Krishan Kumar, and Mohit Gupta. Music generation using three-layered lstm, 2021.

[5] Nikhil Kotecha and Paul Young. Generating music using an lstm network, 2018.

[6] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

CS229 Spring 2023 Final Project